1   SELECTIVE SAMPLING FOR SOUND SIGNAL CLASSIFICATION

2

3   BACKGROUND OF THE INVENTION

4   1.   Field of the Invention

5   The present invention relates generally to systems and methods for sound

6   signal classification, and more particularly to selective sampling techniques for sound

7   signal classification.

8   2.   Discussion of Background Art

9   Interactive Voice Response (IVR) systems are an increasingly important tool

10   for providing information and services in a more cost efficient manner. IVR systems

11   are typically hosted by a server, which includes an array of Digital Signal Processors

12   (DSPs), and enable speakers to interact with corporate databases and services over a

13   telephone using a combination of voice utterances and telephone button presses. IVR

14   systems are particularly cost effective when a large number of speakers require data or

15   services that are very similar in nature and thus can be handled in an automated

16   manner. A speaker using an IVR system may or may not eventually be connected to a

17   live operator, depending upon the complexity of the speaker's request.

18   Due to the significant cost savings often realized with IVR systems, there is a

19   growing demand for such systems to provide more functionality and a richer speaker

20   experience. Toward those ends, IVR systems responsive to a speaker's age range,

21   gender, language, accent, dialect, identity, and so on are desirable. Such functionality

22   often is possible when a speaker's vocal utterance (a.k.a. speech or sound signal) is

23   first digitized and then analyzed, so that a set of meta-data (e.g. the speaker's age

24   range, and so on) can be extracted from the utterance, without requiring the speaker to

25   provide such information directly to the IVR system.

1   While such meta-data extraction has a potential to improve speech recognition

2   of the speaker and enable some novel IVR applications directed to a speaker's

3   particular characteristics, current techniques for meta-data extraction are very

4   computationally intensive and have further burdened IVR system servers and support

5   hardware to the point of creating speed bottlenecks even during normal use.

6   What is needed is a system and method for sound signal classification that

7   overcomes the problems of the prior art.

8

1 <u>SUMMARY OF THE INVENTION</u>

2 The present invention is a system and method for sound signal classification.

3 The method of the present invention includes the elements of: receiving a sound

4 signal; specifying meta-data to be extracted from the sound signal; dividing the sound

5 signal into a set of frames; applying a fitness function to the frames to create a set of

6 fitness data; selecting a frame from the set of frames, if the frame's corresponding

7 fitness datum within the set of fitness data exceeds a predetermined threshold value;

8 extracting the meta-data from the selected frames; and classifying the sound signal

9 based on the meta-data extracted from the selected frames. The system of the present

10 invention includes means for implementing the method.

11 These and other aspects of the invention will be recognized by those skilled in

12 the art upon review of the detailed description, drawings, and claims set forth below.

13

1   <u>BRIEF DESCRIPTION OF THE DRAWINGS</u>

2   Figure 1 is a dataflow diagram of one embodiment of a system for sound

3   signal classification;

4   Figure 2 is one example of a data structure for maintaining a set of fitness data;

5   and

6   Figures 3A and 3B are a flowchart of one embodiment of a method for sound

7   signal classification.

8

1   <u>DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT</u>

2   The present invention discusses a selective frame sampling technique for

3   extracting, from a speaker's voice utterance/speech/sound signal, meta-data such as

4   age range, gender, language, accent, dialect, and identity. The invention not only

5   increases the speed at which an Interactive Voice Response (IVR) system can extract

6   meta-data from (i.e. classify) a speaker's sound signal, but also the accuracy of the

7   extracted meta-data.

8

9   Figure 1 is a dataflow diagram of one embodiment of a system 100 for sound

10   signal classification. Figure 2 is one example of a data structure 200 for maintaining a

11   set of fitness data. Figure 3A and 3B are a flowchart of one embodiment of a method

12   300 for sound signal classification. Figures 1, 2, 3A, and 3B are now discussed

13   together.

14   In step 302, an IVR system 102 receives a vocal utterance/speech/sound signal

15   104 from a speaker. The sound signal 104 will most commonly include of human

16   utterances, such as words, phrases, and sentences. However, the sound signal 104

17   may also include sounds made from a recording, an animal, an inanimate object, and a

18   computer synthesizer. In step 304 the IVR system 102 sends a meta-data request to a

19   classifier selection module 106. The meta-data request specifies what classes of meta-

20   data shall be extracted from the sound signal 104 for a speaker who authored the

21   sound signal 104. The meta-data classes include: age range, gender, language, accent,

22   dialect, identity, and so on. Those skilled in the art, however, will recognize that

23   different types of meta-data may be extracted from the non-human speech/sound

24   signals listed above.

25   In step 306, the IVR system 102 passes the sound signal 104 to a sound signal

26   digitization module 108. The digitization module 108 includes one or more Digital

1   Signal Processors for converting analog sound signals into a digitized form and

2   performing additional processing on the sound signal 104 if necessary. The additional

3   processing may include sound signal noise reduction, echo cancellation, speech

4   detection, and so on. In step 308, the digitization module 108 passes a digitized

5   version of the sound signal 104 back to the IVR system 102 for further processing or

6   storage, depending upon how the IVR system 102 is designed. In step 310, the

7   digitization module 108 passes a digitized version of the sound signal 104 on to a

8   sound signal framing module 110. In step 312, the framing module 110 divides the

9   sound signal 104 into time frames of a predetermined length. Preferably the time

10   frames are of equal length.

11   In step 314, the classifier selection module 106 selects one or more fitness

12   functions, from a fitness function database 112, corresponding to the meta-data to be

13   extracted from the sound signal 104. A fitness function is herein defined as a

14   mathematical calculation to be performed on one or more sound signal frames.

15   While a different fitness function may be used for each class of meta-data to be

16   resolved, preferably a single fitness function that calculates each frame's overall

17   sound signal strength is used. The sound signal strength of a frame is herein defined

18   in the alternative to be: the frame's total signal power, an average of peak amplitudes

19   within the frame, a total energy within the frame, a frame volume equal to a

20   logarithmic value of the sound signal's amplitude, and so on, depending upon a

21   particular implementation of the present invention.

22   In step 316, the classifier selection module 106 passes the selected fitness

23   functions to a frame selection module 114. In step 318, the frame selection module

24   114 applies each of the selected fitness functions individually to each frame received

25   from the sound signal framing module 110, thereby creating the set of fitness data

6

1   200. In step 320, the frame selection module 114 stores the fitness data in a fitness

2   data database 116.

3   Figure 2 shows one example of the data structure 200 for maintaining the set

4   of fitness data. In the example data structure 200, the sound signal 104 has been

5   divided into ten separate frames, labeled by "frame number." A set of fitness data is

6   calculated for each of the meta-data classes (e.g. age range, gender, accent, etc.). An

7   exemplary set of fitness data for meta-data class #1 is shown, and will be used to

8   illustrate the method steps that follow.

9   Generally, the frame selection module 114 preferably identifies a sub-set of the

10   sound signal frames from which the selected meta-data may be accurately extracted.

11   The preferred method selects those sound signal frames that have a greatest relative

12   signal strength for further meta-data extraction.

13   Specifically, in step 322, the frame selection module 114 identifies a greatest

14   fitness datum within a meta-data class (e.g. Frame 5 having a value of 12.0 in the

15   example). In step 324, the frame selection module 114 accesses a predetermined

16   margin (e.g. a margin of 2 in the example) for that meta-data class from the classifier

17   selection module 106. This "margin" effectively sets a sampling rate for the meta-

18   data class. In step 326, the frame selection module 114 calculates a fitness datum

19   threshold equal to the greatest fitness datum minus the margin (e.g. 12.0 minus 2 =

20   10.0 in the example).

21   In step 328, the frame selection module 114 stores a copy of each digitized

22   sound signal frame that has a signal strength equal to or greater than the fitness data

23   threshold (e.g. Frames 4 and 5 in the example) in a sampled frames database 118. In

24   step 330, the frame selection module 114 stores a copy of each digitized sound signal

25   frame that has a signal strength less than the fitness data threshold (e.g. Frames 1-3

26   and 6-10 in the example) in a discarded frames database 120. Alternatively, the

1   frame selection module 114 could just delete these discarded frames. Typically a

2   sound signal's middle frames have a higher Signal-to-Noise Ratio (SNR) (i.e. signal

3   strength) when compared with the sound signal's leading and trailing frames, and thus

4   most often become the sampled frames stored in the sampled frames database 118.

5   As a quick second example, if the margin was set to 4, then the threshold would be

6   12.0 minus 4 = 8, and Frames 1, 3, 4, and 5 would have been stored in the sampled

7   frames database 118.

8

9   In step 332, a classifier module 122 classifies each frame stored in the sampled

10   frames database 118 according to the selected meta-data criteria. In one embodiment

11   of the present invention, the classifier 122 uses a Multi-Layer Perceptron (MLP)

12   neural network trained to recognize the meta-data class patterns.

13   If the sound signal 104 is a speech signal, the MLP neural network will

14   typically have at least three layers: an input layer with 12 nodes, corresponding to the

15   12 Mel-Cepstral components of a speech signal; a hidden layer with 20 nodes; and an

16   output layer with a number of nodes corresponding to each class within the meta-data

17   class (e.g. 2 nodes, "male" and "female," if the meta-data class is "gender"). Back

18   propagation (BP) is used to train the neural network. After being trained on a ground-

19   truth set of about 200,000 frames, the classifier 122 can achieve a meta-data class

20   recognition rate of about 70% for a gender meta-data class at the frame level.

21   Next, in step 334, after having classified each of the sound signals' 104

22   sampled frames individually, the classifier module 122 classifies the entire sound

23   signal 104 according to the selected meta-data classes and stores the result in a sound

24   signal meta-data database 124.

1   One way to classify the entire sound signal 104 is by voting. Voting classifies

2   the sound signal 104 based on which meta-data class is supported by a greatest

3   number of the sampled frames.

4   However, a preferred method for classifying the entire sound signal 104 adds

5   together each of the sampled frame's confidence scores, which were generated by the

6   neural network. That meta-data class with a highest overall total confidence score is

7   chosen as the final class for the entire sound signal 104. The confidence score

8   approach results in a lower classification "error rate," and is even more effective as the

9   "selective sampling" rate is decreased.

10   Another approach classifies the entire sound signal 104 as that class having a

11   statistically longest run-length. The run length of a class is equal to a longest number

12   of continuous sampled frames having been assigned a same meta-data class.

13   Then, in step 336, the sound signal meta-data 124 for the sound signal 104 is

14   provided to the IVR system 102. IVR systems can benefit from such meta-data in a

15   variety of ways, including: improved customer service; added IVR system

16   functionality; and improved statistical record keeping.

17

18   Empirical tests comparing the present invention's selective sampling to even

19   sampling were run on about 1,200 speech files in a 'Test" directory of TIDIGITS

20   corpus. There were about 250,000 frames in total. "Even sampling" is herein defined

21   as when the sampled frames from a sound signal are equally spaced with respect to

22   each other, independent of their signal strength. For example, if the "sampling rate" is

23   1:3, then "even sampling" will select Frames 1,4,7, and 10 in Figure 2, while

24   "selective sampling" will select Frames 1, 3, 4, and 5.

25   It was found that regardless of the "sampling rate", the present invention's

26   "selective sampling" achieved a lower sound signal meta-data 124 error rate when

1   compared to "even sampling." It was also found that there was general "sweet

2   sampling rate spot "for "selective sampling" between 1:2 and 1:3, for gender meta-

3   data classes.   Sampling rates greater than 1:2 tended to include too many low quality

4   sound signal frames, while sampling rates lower than 1:3 tended to discard too many

5   high quality sound signal frames. Those frames discarded by selective sampling also

6   tend to have a lower confidence score than the sampled sound signal frames. Also,

7   since not all of the sound signal's frames are analyzed by the classifier module 122,

8   the speed with which the sound signal meta-data 124 is calculated is also increased.

9

10   While one or more embodiments of the present invention have been described,

11   those skilled in the art will recognize that various modifications may be made.

12   Variations upon and modifications to these embodiments are provided by the present

13   invention, which is limited only by the following claims.